

Starten wir!

Wie bekommen wir Textdaten?

Package und Daten laden

Basisfunktionen in **stringr**

Funktionen mit Regular Expression

Datensätze bändigen & visualisieren mit

B. Philipp Kleer

11. Oktober 2021

Textdaten bearbeiten mit **stringr**

Starten wir!

In diesen Slides wollen wir uns mit Textdaten beschäftigen. Hieru lernen wir grundlegende Funktionen im Umgang mit Textdaten kennen. Eins vorweg: Textdaten sind immer schwieriger zu händeln als numerische Daten. Hierzu werden im späteren Teil auch **regular expressions** eingeführt. Dies bleibt aber auf einem Grundlevel, weiterführend wird dazu sicherlich im Workshop zur quantitativen Textanalyse eingegangen.

In R gibt es innerhalb des **tidyverse** ein spezielles Paket zur Bearbeitung von Textdaten: **stringr**. Gemeinsam mit *regular expressions* können viele Transformationen und Bereinigungen durchgeführt werden.

Dieser Teil des Workshops bedarf aber einer erhöhten Lernbereitschaft (im Vergleich zu anderen Paketen und Funktionen).

Das **Cheat-Sheet**

(<https://github.com/rstudio/cheatsheets/raw/master/strings.pdf>) zu

`stringr` gibt es **hier**

(<https://github.com/rstudio/cheatsheets/raw/master/strings.pdf>).



stringr

Wie bekommen wir Textdaten?

Als mögliche Textdaten fallen einem vermutlich sofort Tweets ein, oder für die Offline-Leute z.B. Wahlprogramme. In anderen Kursen der MethodenTage wird die Erhebung von Social Media-Daten über die sogenannte **API** (*Application Programming Interference*) erklärt. Wir überspringen jetzt diesen Schritt, da es hier erstmal nur um die

grundlegenden Funktionen der Bearbeitung von Textvariablen in R geht.

Dafür habe ich die letzten Tweets der hervorragenden Sendung **The Great British Bake Off** (<https://thegreatbritishbakeoff.co.uk>) über die Twitter-API geladen (Stand: 22. September 2021). Wir werden im nachfolgenden verschiedene Schritte durchführen, wie wir Textdaten mit den Funktionen des Pakets `stringr` bearbeiten können.

Package und Daten laden

Zuerst laden wir wieder `tidyverse` bzw. installieren es, wenn es noch nicht installiert ist:

```
# install.packages("tidyverse")
library("tidyverse")

# alternativ:
# install.packages("stringr")
# library("stringr")
```

Anschließend laden wir den Datensatz `gbbo` ins *environment*.

```
gbbo ← readRDS("../datasets/gbbo.rds")
# oder eigenen Pfad, wenn nicht in Cloud
```

Schauen wir uns nun den Datensatz mal an:

gbbo

user_id <chr>	status_id <chr>	created_at <dtm>
234428224	1440645385249193990	2021-09-22 11:53:20
234428224	1440641707347902473	2021-09-22 11:38:43
234428224	1440630479250620417	2021-09-22 10:54:06
234428224	1440616872618119168	2021-09-22 10:00:02
234428224	1440609316889894912	2021-09-22 09:30:01
234428224	1440601766047928320	2021-09-22 09:00:00
234428224	1440419311588306950	2021-09-21 20:55:00
234428224	1440417068294103042	2021-09-21 20:46:05
234428224	1440416400284139521	2021-09-21 20:43:26
234428224	1440415846640193536	2021-09-21 20:41:14
234428224	1440415353088004103	2021-09-21 20:39:16
234428224	1440414755944947716	2021-09-21 20:36:54
234428224	1440413907982508034	2021-09-21 20:33:31
234428224	1440413834703826955	2021-09-21 20:33:14
234428224	1440413769692168196	2021-09-21 20:32:58

1-15 of 862 rows | 1-3 of 90 columns

Previous12345...58Next

Der Datensatz umfasst in Gänze 90 Variablen, für uns ist in diesem Fall aber nur eine einzige Variable erstmal relevant - nämlich `text`. In `text` befindet sich der Text des jeweiligen Tweets.

Der einfachheit halber speichern wir diese eine Variable als einzelnen Vektor, schaffen eine neue ID-Variable und erstellen den Datensatz `tweet`.

```
tweet <- gbbo %>%
  select(text)

tweet$id <- seq(1,
               862,
               1
               )

head(tweet,
      n = 100
      )
```

text
<chr>
There are two key stages in life. Before you discover Malt Loaf and everything afterwards. #GBBO https://t.co/4LSBTfhiR4
@GeekEnby We all deserve a little ray of sunshine at least once a week. 🙏 #GBBO
Well, it looks like malt loaf awareness isn't in bad shape at all! #GBBO https://t.co/4LSBTfhiR4
We bow down before all of you wonderful people who joined us last night for the return of #GBBO https://t.co/q5ebtUTMAi
When your new crush starts ghosting you... #GBBO https://t.co/XwcVnIBfCi
About last night... #GBBO https://t.co/GulcsAdnJM
Thanks for tweeting along with us tonight on the return of The Great British Bake Off! You really do make the #GBBO Twitter experience pure joy. ❤️❤️ #GBBO https://t.co/PnlzGRb0Mb
Let's get ready to crumble! It's crunch time next Tuesday as Biscuit Week hits the Bake Off Tent! #GBBO https://t.co/xGMRw30WA8
@philringsell @PrueLeith Tres Bon-o. #GBBO https://t.co/vGlYoVXeCs
@Shambom65 True. Absolutely true. #GBBO
Disclaimer: This is not a percentage in terms of flavour balance. It's how many people like Mint Choc Ch... *stares directly to camera* #GBBO
It's official! Mint Choc Chip has now been certified as being 67.6% Mint. #GBBO https://t.co/vUovHq8iqD
"Although my time in the Bake Off tent may have been short, it certainly has been very sweet." \nHere's Tom's lovely letter to all you #GBBO fans. https://t.co/JgKlfD8WDg
No expense was spared to turn the Bake Off Tent into an anti-gravity zone for tonight's show. #GBBO https://t.co/8MEF5qemLe
Mark our words... 🤔 #GBBO https://t.co/5s8w8XjVDR
1-15 of 100 rows 1-1 of 2 columns

Previous 1 2 3 4 5 ... 7 Next

Basisfunktionen in stringr

Die Basisfunktionen, die wir jetzt zu Beginn kennenlernen, sind folgende ...

- `str_length()`
- `str_sub()`
- `str_dup()`
- `str_pad()`
- `str_trunc()`

- `str_trim()`
- `str_to_upper()` / `str_to_lower()` / `str_to_title()`

str_length()

Die Funktion ist ähnlich zur Funktion `nchar()` aus `baseR`. Sie gibt die Länge der jeweiligen Zeichenkette an. Wir können dies für alle ausgeben lassen oder für einzelne Tweets. Wenn wir es für alle ausgeben lassen, ist die Ausgabe wieder ein Vektor, aber eben numerisch.

```
str_length(tweet$text)
```

```

[1] 120 79 96 119 70 47 179 127 66 39 141 103 168 123 49 173 112 146
[19] 86 97 91 90 96 90 76 95 83 96 111 40 108 86 86 81 67 108
[37] 88 96 102 33 139 24 16 50 104 38 190 61 85 37 31 46 63 80
[55] 95 40 96 90 40 76 59 58 119 182 108 33 119 103 73 105 67 80
[73] 86 50 53 96 136 30 80 73 83 77 59 36 47 60 42 78 87 49
[91] 149 181 59 66 48 55 51 55 116 78 119 91 112 69 26 178 43 78
[109] 81 81 81 118 56 259 143 66 83 290 158 83 93 99 125 54 125 112
[127] 58 118 40 296 121 84 56 121 67 28 58 106 67 26 83 61 88 106
[145] 158 82 132 45 85 21 160 49 116 84 55 123 130 186 99 52 40 49
[163] 138 54 59 112 64 125 245 69 155 141 170 164 161 151 122 148 122 114
[181] 112 146 121 87 171 173 202 195 71 93 58 245 192 230 124 152 42 29
[199] 119 173 123 120 108 33 139 111 41 141 122 119 53 72 40 50 74 150
[217] 107 38 90 40 56 61 63 129 59 91 67 113 78 99 99 41 120 174
[235] 100 22 92 108 48 89 39 51 140 64 118 152 173 116 180 122 87 287
[253] 38 68 108 206 183 152 26 102 35 50 64 70 159 129 58 66 66 87
[271] 98 89 253 100 77 41 46 181 62 121 74 67 136 31 87 71 147 89
[289] 57 99 55 45 83 107 170 136 281 83 115 75 76 113 94 128 84 40
[307] 56 111 63 80 147 155 126 77 63 73 61 39 143 164 81 115 73 67
[325] 97 26 36 170 80 49 139 20 71 67 39 81 112 129 161 54 101 103
[343] 148 139 84 89 72 67 147 158 68 155 131 155 108 109 40 90 119 43
[361] 59 46 37 67 72 112 72 70 115 43 76 61 102 130 178 118 196 164
[379] 85 92 97 156 44 87 62 293 57 51 71 139 152 169 40 155 148 102
[397] 155 91 158 127 42 39 41 68 98 61 86 86 119 48 66 36 102 110
[415] 72 79 118 46 54 85 98 99 112 58 89 70 50 122 84 108 162 21
[433] 68 42 131 62 164 92 196 111 92 153 148 117 100 43 126 49 68 154
[451] 154 125 150 38 151 104 145 56 126 132 69 59 96 59 52 56 70 83
[469] 40 72 78 91 36 114 93 98 36 143 40 85 46 121 69 74 34 108
[487] 52 125 65 143 188 89 66 157 171 93 230 50 65 33 112 38 153 48
[505] 103 147 82 105 152 58 36 166 189 182 144 41 98 62 88 136 59 64
[523] 42 44 45 41 84 140 31 142 45 93 82 105 143 87 70 40 85 106
[541] 93 47 130 98 155 198 76 155 176 84 80 51 140 109 138 72 74 91
[559] 115 111 41 152 118 151 94 215 147 49 71 65 141 128 81 46 72 196
[577] 49 139 76 112 136 73 40 52 79 62 93 86 99 58 56 79 107 139
[595] 106 125 278 235 211 180 125 84 61 77 105 38 79 70 92 167 142 172
[613] 73 138 160 138 197 41 92 41 149 47 99 79 76 76 39 45 50 120
[631] 50 99 76 68 86 115 38 15 71 58 64 41 132 81 33 40 152 38
[649] 53 59 52 127 125 120 181 155 155 87 89 82 63 40 74 90 87 147
[667] 149 74 73 42 43 124 293 144 129 58 31 161 126 103 67 43 101 127
[685] 96 68 120 40 61 109 88 102 46 113 71 64 112 59 38 130 98 102
[703] 42 87 88 55 59 75 135 45 59 28 109 85 41 91 72 95 50 64
[721] 151 116 128 96 45 154 94 205 278 174 189 175 199 130 157 282 290 284
[739] 257 294 194 46 77 173 236 72 47 163 261 84 268 171 157 57 85 42
[757] 109 134 137 147 108 89 188 40 168 71 65 100 149 89 228 279 104 91
[775] 96 78 64 61 124 104 43 125 59 239 45 34 100 85 115 77 86 88
[793] 33 77 45 60 78 112 220 200 60 119 58 250 209 74 239 129 257 208
[811] 197 241 201 237 83 102 42 61 144 178 68 181 179 219 114 226 116 130
[829] 167 101 54 151 82 169 154 161 152 81 238 77 67 58 100 78 141 159
[847] 118 43 51 255 289 86 49 41 68 42 76 86 108 103 41 79

```

```

# oder für einen spezifischen
str_length(tweet$text[23])

```

```

[1] 96

```

```
# oder mit piping
tweet %>%
  filter(id == 275) %>%
  select(text) %>%
  str_length()
```

```
[1] 77
```

str_sub()

Wir können uns auch einzelne Zeichen aus einem String ausgeben lassen. So z.B. das zweite bis fünfte Zeichen aus dem 23. Tweet.

```
# hier der 23. tweet
str_sub(tweet$text[23],
        2,
        5
        )
```

```
[1] "iuse"
```

```
tweet %>%
  filter(id == 23) %>%
  select(text) %>%
  str_sub(2,
        5
        )
```

```
[1] "iuse"
```

Natürlich kann man sich auch einzelne Buchstaben ausgeben lassen, dazu setzt man den Anfang (zweite Argument) und das Ende der Extraktion (dritte Argument) gleich. Will man von hinten zählen nimmt man einfach ein - zur Hilfe.

```
# hier der 23. tweet
str_sub(tweet$text[23],
        25,
        25
        )
```

```
[1] " "
```

```
str_sub(tweet$text[23],
        -23,
        -2
        )
```

```
[1] "https://t.co/WxwxZquxP"
```

Eine kleine Denkaufgabe: Warum ist die erste Ausgabe leer?

Alternativ kann man so auch bestimmte Stellen im Text ersetzen, sofern man deren numerische Stellen kennt. Wie dies einfacher mit *regular expression* geht, lernen wir später.

```
tweet$text[23]
```

```
[1] "Giuseppe's 'Jack and The Beanstalk' Anti-Gravity Showstopper Cake. #GBBO https://t.co/WxwxZquxPi"
```

```
str_sub(tweet$text[23],  
        26,  
        26  
        ) ← "M"
```

```
tweet$text[23]
```

```
[1] "Giuseppe's 'Jack and The Meanstalk' Anti-Gravity Showstopper Cake. #GBBO https://t.co/WxwxZquxPi"
```

str_dup()

Anstatt Zeichen zu löschen oder zu ersetzen, können wir Zeichen auch duplizieren. Wiederholen wir einfach den Text von Tweet 23 viermal.

```
str_dup(tweet$text[23],  
        4  
        )
```

```
[1] "Giuseppe's 'Jack and The Meanstalk' Anti-Gravity Showstopper Cake. #GBBO https://t.co/WxwxZquxPiGiuseppe's 'Jack and The Meanstalk' Anti-Gravity Showstopper Cake. #GBBO https://t.co/WxwxZquxPiGiuseppe's 'Jack and The Meanstalk' Anti-Gravity Showstopper Cake. #GBBO https://t.co/WxwxZquxPiGiuseppe's 'Jack and The Meanstalk' Anti-Gravity Showstopper Cake. #GBBO https://t.co/WxwxZquxPi"
```

Man kann das ganze auch verketten und wieder direkt *pipen* mit `tidyverse`.

```
tweet %>%  
  filter(id = 23) %>%  
  select(text) %>%  
  str_sub(1,  
        8  
        ) %>%  
  str_dup(4)
```

```
[1] "GiuseppeGiuseppeGiuseppeGiuseppe"
```

Kleine Denkaufgabe: Was wird hier wiederholt?

str_pad()

Oft ist es von Vorteil, wenn Zeichenketten dieselbe Länge aufweisen, z.B. wenn man sie gleichmäßig grafisch darstellen möchte. Dies können wir mit `str_pad()` machen: Die Zeichenkette soll 100 Zeichen lang werden. Zuerst erstellen wir dazu aber einfach einen Character-Vektor, der nur zwei Tweets beinhaltet (Nr. 23 und 731).

```
partOfTweet <- tweet$text[c(23,
                             731
                             )
                             ]
```

partOfTweet

```
[1] "Giuseppe's 'Jack and The Meanstalk' Anti-Gravity Showstopper Cake. #GBBO https://t.co/WxwxZquxPi"
[2] "It's got to be perfect. Join us tonight for a new series of Bake Off: The Professionals with @BenoitBlin_MCA, @LiamcBakes, @TomAllenComedy and @Cherish_Finden! #GBBO https://t.co/PamvTHcRsv"
```

Wir haben jetzt also einen einzelnen Vektor, der zwei Tweets beinhaltet. Jetzt wollen wir die Länge auf 100 festsetzen:

```
partOfTweet <- str_pad(partOfTweet,
                        100,
                        side = "both"
                        )
```

partOfTweet

```
[1] " Giuseppe's 'Jack and The Meanstalk' Anti-Gravity Showstopper Cake. #GBBO https://t.co/WxwxZquxPi "
[2] "It's got to be perfect. Join us tonight for a new series of Bake Off: The Professionals with @BenoitBlin_MCA, @LiamcBakes, @TomAllenComedy and @Cherish_Finden! #GBBO https://t.co/PamvTHcRsv"
```

Kleine Denkaufgabe: Was passiert hier nun beim zweiten Tweet?

Anscheinend nichts. Und das ist auch gut so! Denn ansonsten hätte die Funktion den Tweet gekürzt und dies macht die Funktion eben nicht. Ist ein Text länger als es in `str_pad()` angegeben ist, bleibt der Tweet so wie er ist, erhalten.

Eleganter Pipen

Oben haben wir ja bereits mit tidyverse **pipes** benutzt, dies machen wir jetzt aber eleganter. Die Funktionen aus `stringr` haben alle kein Datenargument, daher sind sie auch nicht direkt zu pipen. Das führt immer zu Fehlermeldungen.

```
tweet <- tweet %>%
  str_pad(text,
          100,
          side = "both"
          )
```

Die Funktionen aus `stringr` sind allesamt direkt Mutationen eines Vektors. Mithilfe von `mutate()` können wir diese aber auch mit **Pipes** benutzen. Wir definieren eine neue Variable und dieser Variable ordnen wir die Mutation über die `stringr`-Funktion zu.


```
tweet <- tweet %>%  
  mutate(shortT = str_pad(text,  
                           100,  
                           side = "both"  
                           )  
  )  
  
head(tweet$shortT,  
      n = 50  
      )
```

[1] "There are two key stages in life. Before you discover Malt Loaf and everything afterwards. #GBBO <https://t.co/4LSBTfhiR4>"

[2] " @GeekEnby We all deserve a little ray of sunshine at least once a week. 🙏 #GBBO "

[3] " Well, it looks like malt loaf awareness isn't in bad shape at all! #GBBO <https://t.co/4LSBTfhiR4> "

[4] "We bow down before all of you wonderful people who joined us last night for the return of #GBBO <https://t.co/q5ebtUTMAi>"

[5] " When your new crush starts ghosting you... #GBBO <https://t.co/XwcVnIBfCi> "

[6] " About last night... #GBBO <https://t.co/GulcsAdnJM> "

[7] "Thanks for tweeting along with us tonight on the return of The Great British Bake Off! You really do make the #GBBO Twitter experience pure joy. ❤️❤️ #GBBO <https://t.co/PnlzGRb0Mb>"

[8] "Let's get ready to crumble! It's crunch time next Tuesday as Biscuit Week hits the Bake Off Tent! #GBBO <https://t.co/xGMRw30WA8>"

[9] " @philringsell @PrueLeith Tres Bon-o. #GBBO <https://t.co/vGlYoVXeCs> "

[10] " @Shambom65 True. Absolutely true. #GBBO "

[11] "Disclaimer: This is not a percentage in terms of flavour balance. It's how many people like Mint Choc Ch... *stares directly to camera* #GBBO"

[12] "It's official! Mint Choc Chip has now been certified as being 67.6% Mint. #GBBO <https://t.co/vUovHq8iqD>"

[13] "“Although my time in the Bake Off tent may have been short, it certainly has been very sweet.”\nHere's Tom's lovely letter to all you #GBBO fans. <https://t.co/JgKlfd8WDg>"

[14] "No expense was spared to turn the Bake Off Tent into an anti-gravity zone for tonight's show. #GBBO <https://t.co/8MEF5qemLe>"

[15] " Mark our words ... 🤔 #GBBO <https://t.co/5s8w8XjVDR> "

[16] "Poor Tom! So gracious in defeat - and we're so sad to see him go. But what time we did have with him in the Tent has been utterly fantastic. ❤️ #GBBO <https://t.co/qrz9FWe9sc>"

[17] "He came. He self-judged. He won. Here's how Star Baker Jürgen conquered Cake Week! #GBBO <https://t.co/kKJFeUvAUk>"

[18] "One episode into Bake Off 2021 - and George is already turning into one of the most relatable bakers we've ever met. #GBBO <https://t.co/pKuPc4f4IZ>"

[19] " When Jürgen's victory call home didn't quite go to plan... #GBBO <https://t.co/RzgdhqFmnb> "

[20] " Do you have cuppa hot chocolate in a medium-sized tinted glass mug? #GBBO <https://t.co/xbWbk4JJnL> "

[21] " A massive thumbs up for Star Baker Jürgen from the judges! 👍👍 #GBBO <https://t.co/TgZUbqcSSb> "

[22] " Jürgen's 'Night Time Reading' Anti-Gravity Showstopper Cake. #GBBO <https://t.co/42rFL35omD> "

[23] " Giuseppe's 'Jack and The Meanstalk' Anti-Gravity Showstopper Cake. #GBBO <https://t.co/WxwxZquxPi> "

[24] " Tom's Classically Spectacular Anti-Gravity Showstopper Cake. #GBBO <https://t.co/JSdm6M3Zt3> "

[25] " Rochica's Anti-Gravity Apple Cake Showstopper. #GBBO <https://t.co/60Vanr8fnB> "

[26] " Crystelle's 'Bouq-cake' of Flowers Anti-Gravity Showstopper Cake. #GBBO <https://t.co/lilsAPVK6n> "

[27] " Freya's 'Flower Potty' Anti-Gravity Showstopper Cake. #GBBO <https://t.co/OFrqBxqiWf> "

```

[28] "    Lizzie's Nana's Peaches & Cream Anti-Gravity Showstopper Cake. #GBBO https://t.co/4t4c1VxxcS "
[29] "FACT: The gravitational pull in the Bake Off Tent is the strongest on the planet. #GBBO https://t.co/LyC3cPKkFG"
[30] "                                @courtsthomasx Congratulations! 🥳🥳 #GBBO
"
[31] "Amanda's Elderflower & Lemon Surf Explosion Anti-Gravity Showstopper Cake. #GBBO https://t.co/6C8pShegFt"
[32] "    Jairzeno's 'Kick The Can' Anti-Gravity Showstopper Cake. #GBBO https://t.co/M2n txYuy8X
"
[33] "    Maggie's Ice Cream Cornet Anti-Gravity Showstopper Cake. #GBBO https://t.co/qyW u2RSC0i
"
[34] "    George's Greek Table Anti-Gravity Showstopper Cake. #GBBO https://t.co/Yt0tSl ogb2
"
[35] "                                This anti-gravity cake had one job... #GBBO https://t.co/pSMD8uILn4
"
[36] "Any Chuckle Brothers homage on primetime television is music to our ears. ❤️❤️ #GBBO https://t.co/ZZmdn49fJQ"
[37] "    Chigs' 'Where It All Began' Anti-Gravity Showstopper Cake. #GBBO https://t.co/1s cORW3clT
"
[38] "    The pure joy of baking something you're proud of. Can't be beaten. #GBBO https://t.co/yB6TTaWcZS "
[39] "Unfortunately Prue2 is already the name of @PrueLeith's U2 tribute band. #GBBO https://t.co/Pm97wOWQNn"
[40] "                                🤔🤔🤔 #GBBO https://t.co/8jug1xBll1
"
[41] "The first rule of Bake Off:\nDo not stare back at Paul\n\nThe second rule of Bake Off:\nDo not stare back at Paul\n#GBBO https://t.co/XbZqAE5mjA"
[42] "                                @psyche1701 Maybe. #GBBO
"
[43] "                                WISDOMOUS! #GBBO
"
[44] "                                @JadeForster A treat for the whole family. 🍰 #GBBO
"
[45] "Anxiously looking at all the new bakers and not wanting any of them to go. #GBBO https://t.co/02YguY0MWm"
[46] "                                Oh dear. #GBBO https://t.co/4FS3X12A0r
"
[47] "1. Mint Choc Chip is the entry level ice cream for 93% of the United Kingdom. But please don't Google that.\n\n2. \"MagPrue\" can only lead to conspiracy theories.\n\n#GBBO https://t.co/95cC26eXOZ"
[48] "                                Please do not try this at home. #GBBO https://t.co/iGwsPuwZBG
"
[49] "    Just to clear up the confusion about Prue and Maggie... #GBBO https://t.co/qmm6 CGIHS6
"
[50] "                                @laurajaycrochet *snaps pencil* #GBBO
"

```

Es ist zwar ein Umweg, aber ein guter: Denn so bleibt die Ursprungsvariable immer im Datensatz erhalten. Nicht zu empfehlen ist es die Originalvariable direkt in `mutate()` zu überschreiben.

str_trunc()

Um gleichzeitig die Tweets zu kürzen, benutzt man die Funktion `str_trunc()`. Hiermit kann die maximale Zeichenkettenzahl gesetzt werden. So könnten wir zum Beispiel sagen, uns interessiert nur der Beginn der Tweets, also die ersten 50 Zeichen.

```
tweet <- tweet %>%
  mutate(truncT = str_trunc(text,
                             50
                           ),
         truncT = str_pad(truncT,
                           60,
                           side = "both"
                          )
        )

head(tweet$truncT,
      n = 50
    )
```

[1] " There are two key stages in life. Before you di... "

[2] " @GeekEnby We all deserve a little ray of sunshi... "

[3] " Well, it looks like malt loaf awareness isn't i... "

[4] " We bow down before all of you wonderful people ... "

[5] " When your new crush starts ghosting you... #GBBO ... "

[6] " About last night... #GBBO <https://t.co/GulcsAdnJM> "

[7] " Thanks for tweeting along with us tonight on th... "

[8] " Let's get ready to crumble! It's crunch time ne... "

[9] " @philringsell @PrueLeith Tres Bon-o. #GBBO [http...](http://t.co/5s8w8XjVDR) "

[10] " @Shambom65 True. Absolutely true. #GBBO "

[11] " Disclaimer: This is not a percentage in terms o... "

[12] " It's official! Mint Choc Chip has now been cert... "

[13] " "Although my time in the Bake Off tent may have... "

[14] " No expense was spared to turn the Bake Off Tent... "

[15] " Mark our words... 🙄 #GBBO <https://t.co/5s8w8XjVDR> "

[16] " Poor Tom! So gracious in defeat - and we're so ... "

[17] " He came. He self-judged. He won. Here's how Sta... "

[18] " One episode into Bake Off 2021 - and George is ... "

[19] " When Jürgen's victory call home didn't quite go ... "

[20] " Do you have cuppa hot chocolate in a medium-siz... "

[21] " A massive thumbs up for Star Baker Jürgen from ... "

[22] " Jürgen's 'Night Time Reading' Anti-Gravity Show... "

[23] " Giuseppe's 'Jack and The Meanstalk' Anti-Gravit... "

[24] " Tom's Classically Spectacular Anti-Gravity Show... "

[25] " Rochica's Anti-Gravity Apple Cake Showstopper. ... "

[26] " Crystelle's 'Bouq-cake' of Flowers Anti-Gravity... "

[27] " Freya's 'Flower Potty' Anti-Gravity Showstopper... "

[28] " Lizzie's Nana's Peaches & Cream Anti-Gravit... "

[29] " FACT: The gravitational pull in the Bake Off Te... "

[30] " @courtstthomasx Congratulations! 🥳🥳 #GBBO "

[31] " Amanda's Elderflower & Lemon Surf Explosion... "

[32] " Jairzeno's 'Kick The Can' Anti-Gravity Showstop... "

[33] " Maggie's Ice Cream Cornet Anti-Gravity Showstop... "

[34] " George's Greek Table Anti-Gravity Showstopper C... "

[35] " This anti-gravity cake had one job... #GBBO [htt...](http://t.co/8jug1xBLL1) "

[36] " Any Chuckle Brothers homage on primetime televi... "

[37] " Chigs' 'Where It All Began' Anti-Gravity Showst... "

[38] " The pure joy of baking something you're proud o... "

[39] " Unfortunately Prue2 is already the name of @Pru... "

[40] " 🙄🙄🙄 #GBBO <https://t.co/8jug1xBLL1> "

[41] " The first rule of Bake Off:\nDo not stare back a... "

[42] " @psyche1701 Maybe. #GBBO "

[43] " WISDOMOUS! #GBBO "

[44] " @JadeForster A treat for the whole family. 🥳 #GBBO "

[45] " Anxiously looking at all the new bakers and not... "

[46] " Oh dear. #GBBO <https://t.co/4FS3X12A0r> "

[47] " 1. Mint Choc Chip is the entry level ice cream ... "

[48] " Please do not try this at home. #GBBO [https://t...](https://t.co/4FS3X12A0r) "

[49] " Just to clear up the confusion about Prue and M... "

[50] " @laurajaycrochet *snaps pencil* #GBBO "

Wie man sieht, werden Tweets die länger als 50 Zeichen sind mit ... am Ende gekennzeichnet.

str_trim()

Hat man am Ende oder Beginn Leerzeichen, nutzt man einfach `str_trim()`, um diese zu entfernen. Die zuvor geschaffenen Leerzeichen, entfernen wir jetzt wieder.

```
tweet <- tweet %>%  
  mutate(truncT2 = str_trim(truncT,  
                             side = "both"  
                           )  
        )  
  
head(tweet$truncT2,  
      n = 50  
    )
```

[1] "There are two key stages in life. Before you di ... "

[2] "@GeekEnby We all deserve a little ray of sunshi ... "

[3] "Well, it looks like malt loaf awareness isn't i ... "

[4] "We bow down before all of you wonderful people ... "

[5] "When your new crush starts ghosting you... #GBBO ... "

[6] "About last night... #GBBO <https://t.co/GulcsAdnJM>"

[7] "Thanks for tweeting along with us tonight on th ... "

[8] "Let's get ready to crumble! It's crunch time ne ... "

[9] "@philringsell @PrueLeith Tres Bon-o. #GBBO http ... "

[10] "@Shambom65 True. Absolutely true. #GBBO"

[11] "Disclaimer: This is not a percentage in terms o ... "

[12] "It's official! Mint Choc Chip has now been cert ... "

[13] ""Although my time in the Bake Off tent may have ... "

[14] "No expense was spared to turn the Bake Off Tent ... "

[15] "Mark our words ... 😊 #GBBO <https://t.co/5s8w8XjVDR>"

[16] "Poor Tom! So gracious in defeat - and we're so ... "

[17] "He came. He self-judged. He won. Here's how Sta ... "

[18] "One episode into Bake Off 2021 - and George is ... "

[19] "When Jürgen's victory call home didn't quite go ... "

[20] "Do you have cuppa hot chocolate in a medium-siz ... "

[21] "A massive thumbs up for Star Baker Jürgen from ... "

[22] "Jürgen's 'Night Time Reading' Anti-Gravity Show ... "

[23] "Giuseppe's 'Jack and The Meanstalk' Anti-Gravit ... "

[24] "Tom's Classically Spectacular Anti-Gravity Show ... "

[25] "Rochica's Anti-Gravity Apple Cake Showstopper. ... "

[26] "Crystelle's 'Bouq-cake' of Flowers Anti-Gravity ... "

[27] "Freya's 'Flower Potty' Anti-Gravity Showstopper ... "

[28] "Lizzie's Nana's Peaches & Cream Anti-Gravit ... "

[29] "FACT: The gravitational pull in the Bake Off Te ... "

[30] "@courtsthomasx Congratulations! 🥳🥳 #GBBO"

[31] "Amanda's Elderflower & Lemon Surf Explosion ... "

[32] "Jairzeno's 'Kick The Can' Anti-Gravity Showstop ... "

[33] "Maggie's Ice Cream Cornet Anti-Gravity Showstop ... "

[34] "George's Greek Table Anti-Gravity Showstopper C ... "

[35] "This anti-gravity cake had one job... #GBBO htt ... "

[36] "Any Chuckle Brothers homage on primetime televi ... "

[37] "Chigs' 'Where It All Began' Anti-Gravity Showst ... "

[38] "The pure joy of baking something you're proud o ... "

[39] "Unfortunately Prue2 is already the name of @Pru ... "

[40] "😭😭😭 #GBBO <https://t.co/8jug1xBLL1>"

[41] "The first rule of Bake Off:\nDo not stare back a ... "

[42] "@psyche1701 Maybe. #GBBO"

[43] "WISDOMOUS! #GBBO"

[44] "@JadeForster A treat for the whole family. 🥰 #GBBO"

[45] "Anxiously looking at all the new bakers and not ... "

[46] "Oh dear. #GBBO <https://t.co/4FS3X12AOr>"

[47] "1. Mint Choc Chip is the entry level ice cream ... "

[48] "Please do not try this at home. #GBBO <https://t.co/4FS3X12AOr>"

[49] "Just to clear up the confusion about Prue and M ... "

[50] "@laurajaycrochet *snaps pencil* #GBBO"

str_to_upper() / str_to_lower() / str_to_title()

Groß- und Kleinschreibung spielt gerade bei der quantitativen Analyse von Texten keine Rolle. Daher ändert man in der Regel alle Zeichen auf Großbuchstaben (`str_to_upper()`) oder auf Kleinbuchstaben (`str_to_lower`). Teste gleich anschließend selbst aus, was die Funktion `str_to_title()` macht!

```
tweet <- tweet %>%  
  mutate(truncT2 = str_to_lower(truncT2  
                                )  
        )  
  
head(tweet$truncT2,  
      n = 50  
      )
```


[1] "there are two key stages in life. before you di ... "

[2] "@geekenby we all deserve a little ray of sunshi ... "

[3] "well, it looks like malt loaf awareness isn't i ... "

[4] "we bow down before all of you wonderful people ... "

[5] "when your new crush starts ghosting you... #gbbo ... "

[6] "about last night... #gbbo <https://t.co/gulcsadnjm>"

[7] "thanks for tweeting along with us tonight on th ... "

[8] "let's get ready to crumble! it's crunch time ne ... "

[9] "@philringsell @prueleith tres bon-o. #gbbo http ... "

[10] "@shambom65 true. absolutely true. #gbbo"

[11] "disclaimer: this is not a percentage in terms o ... "

[12] "it's official! mint choc chip has now been cert ... "

[13] ""although my time in the bake off tent may have ... "

[14] "no expense was spared to turn the bake off tent ... "

[15] "mark our words ... 🙄 #gbbo <https://t.co/5s8w8xjvdr>"

[16] "poor tom! so gracious in defeat - and we're so ... "

[17] "he came. he self-judged. he won. here's how sta ... "

[18] "one episode into bake off 2021 - and george is ... "

[19] "when jürgen's victory call home didn't quite go ... "

[20] "do you have cuppa hot chocolate in a medium-siz ... "

[21] "a massive thumbs up for star baker jürgen from ... "

[22] "jürgen's 'night time reading' anti-gravity show ... "

[23] "giuseppe's 'jack and the meanstalk' anti-gravit ... "

[24] "tom's classically spectacular anti-gravity show ... "

[25] "rochica's anti-gravity apple cake showstopper. ... "

[26] "crystelle's 'bouq-cake' of flowers anti-gravity ... "

[27] "freya's 'flower potty' anti-gravity showstopper ... "

[28] "lizzie's nana's peaches & cream anti-gravit ... "

[29] "fact: the gravitational pull in the bake off te ... "

[30] "@courtsthomasx congratulations! 🥳🥳 #gbbo"

[31] "amanda's elderflower & lemon surf explosion ... "

[32] "jairzeno's 'kick the can' anti-gravity showstop ... "

[33] "maggie's ice cream cornet anti-gravity showstop ... "

[34] "george's greek table anti-gravity showstopper c ... "

[35] "this anti-gravity cake had one job... #gbbo htt ... "

[36] "any chuckle brothers homage on primetime televi ... "

[37] "chigs' 'where it all began' anti-gravity showst ... "

[38] "the pure joy of baking something you're proud o ... "

[39] "unfortunately prue2 is already the name of @pru ... "

[40] "😭😭😭 #gbbo <https://t.co/8jug1xbll1>"

[41] "the first rule of bake off:\ndo not stare back a ... "

[42] "@psyche1701 maybe. #gbbo"

[43] "wisdomous! #gbbo"

[44] "@jadeforster a treat for the whole family. 🥰 #gbbo"

[45] "anxiously looking at all the new bakers and not ... "

[46] "oh dear. #gbbo <https://t.co/4fs3x12aor>"

[47] "1. mint choc chip is the entry level ice cream ... "

[48] "please do not try this at home. #gbbo <https://t.co/4fs3x12aor>"

[49] "just to clear up the confusion about prue and m ... "

[50] "@laurajaycrochet *snaps pencil* #gbbo"

Das war es zu den Basis-Funktionen. Jetzt gehen wir über zu Funktionen, die insbesondere in Kombination mit **regular expressions** helfen, Textdaten für die Analyse gut vorzubereiten.

Funktionen mit Regular Expression

Mit **regular expression** (*regex*) können Zeichenkette bestimmt werden, die beliebig eingesetzt werden können. Dabei werden per Code gewisse Muster vorgegeben, mithilfe denen R diese im Text erkennt und die gewünschte Transformation vornimmt. Es ist eine spezifische Art mit Textdaten umzugehen und zu identifizieren.

Eine gute Übersicht gibt es auf **Wikipedia** (https://de.wikipedia.org/wiki/Regulärer_Ausdruck#Syntax) oder im **Cheat-Sheet von stringr** (<https://github.com/rstudio/cheatsheets/raw/master/strings.pdf>)

Hier der Ausschnitt aus dem **Cheat-Sheet**.


Regular Expressions -

Regular expressions, or *regexps*, are a concise language for describing patterns in strings.

see <- function(rx) str_view_all("abc ABC 123\t.!?\\()\n", rx)

string (type this)	regex (to mean this)	matches (which matches this)	example
a (etc.)	a	a (etc.)	see("a")
.	.	.	see(".", "abc ABC 123")
!	!	!	see("!", "abc ABC 123")
?	?	?	see("?", "abc ABC 123")
\\	\\	\\	see("\\", "abc ABC 123")
(((see("(", "abc ABC 123")
)))	see(")", "abc ABC 123")
{	{	{	see("{", "abc ABC 123")
}	}	}	see("}", "abc ABC 123")
\\n	\\n	new line (return)	see("\\n", "abc ABC 123")
\\t	\\t	tab	see("\\t", "abc ABC 123")
\\s	\\s	any whitespace (S for non-whitespaces)	see("\\s", "abc ABC 123")
\\d	\\d	any digit (D for non-digits)	see("\\d", "abc ABC 123")
\\w	\\w	any word character (W for non-word chars)	see("\\w", "abc ABC 123")
\\b	\\b	word boundaries	see("\\b", "abc ABC 123")
[digit:]	[digit:]	digits	see("[digit:]", "abc ABC 123")
[alpha:]	[alpha:]	letters	see("[alpha:]", "abc ABC 123")
[lower:]	[lower:]	lowercase letters	see("[lower:]", "abc ABC 123")
[upper:]	[upper:]	uppercase letters	see("[upper:]", "abc ABC 123")
[alnum:]	[alnum:]	letters and numbers	see("[alnum:]", "abc ABC 123")
[punct:]	[punct:]	punctuation	see("[punct:]", "abc ABC 123")
[graph:]	[graph:]	letters, numbers, and punctuation	see("[graph:]", "abc ABC 123")
[space:]	[space:]	space characters (i.e. \\s)	see("[space:]", "abc ABC 123")
[blank:]	[blank:]	space and tab (but not new line)	see("[blank:]", "abc ABC 123")
.	.	every character except a new line	see(".", "abc ABC 123")

¹ Many base R functions require classes to be wrapped in a second set of [], e.g. **[[:digit:]]**



[space:]
new line

[blank:]
space
tab

[graph:]

[punct:]
, . : ; ? ! / * @ #
- _ " ' [] { } ()

[symbol:]
| ` = + ^
~ < > \$

[alnum:]
[digit:]
0 1 2 3 4 5 6 7 8 9

[alpha:]

[lower:]
a b c d e f
g h i j k l
m n o p q r
s t u v w x
y z

[upper:]
A B C D E F
G H I J K L
M N O P Q R
S T U V W X
Y Z

ALTERNATES

alt <- function(rx) str_view_all("abcde", rx)

regex	matches	example
ab d	or	alt("ab d", "abcde")
[abe]	one of	alt("[abe]", "abcde")
[^abe]	anything but	alt("[^abe]", "abcde")
[a-c]	range	alt("[a-c]", "abcde")

QUANTIFIERS

quant <- function(rx) str_view_all("a.aa.aaa", rx)

regex	matches	example
a?	zero or one	quant("a?", "a.aa.aaa")
a*	zero or more	quant("a*", "a.aa.aaa")
a+	one or more	quant("a+", "a.aa.aaa")
a{n}	exactly n	quant("a{2}", "a.aa.aaa")
a{n,}	n or more	quant("a{2,}", "a.aa.aaa")
a{n,m}	between n and m	quant("a{2,4}", "a.aa.aaa")

ANCHORS

anchor <- function(rx) str_view_all("aaa", rx)

regex	matches	example
^a	start of string	anchor("^a", "aaa")
a\$	end of string	anchor("a\$", "aaa")

GROUPS

ref <- function(rx) str_view_all("abbaab", rx)

Use parentheses to set precedent (order of evaluation) and create groups

regex	matches	example
(ab)d(e)	sets precedence	alt("(ab)d(e)", "abcde")

Use an escaped number to refer to and duplicate parentheses groups that occur earlier in a pattern. Refer to each group by its order of appearance

string (type this)	regex (to mean this)	matches (which matches this)	example (the result is the same as ref("abba"))
\\1	\\1 (etc.)	first () group, etc.	ref("(a)(b)\\2\\1")

LOOK AROUNDS

look <- function(rx) str_view_all("bacad", rx)

regex	matches	example
a(?=c)	followed by	look("a(?=c)", "bacad")
a(?!c)	not followed by	look("a(?!c)", "bacad")
(?<=b)a	preceded by	look("(?<=b)a", "bacad")
(?<!b)a	not preceded by	look("(?<!b)a", "bacad")

Regular expressions in R

Die folgenden Funktionen werden oftmals in Kombination mit *regex* eingesetzt, um String-Variablen zu bearbeiten und wir werden diese im Folgenden genauer kennenlernen.

- `str_detect()` / `str_count()`

- `str_extract()`
- `str_subset()`
- `str_replace()`

`str_detect()` / `str_count()`

Machen wir der Einfachheit halber nochmal ein Subsample, dass nur 50 Tweets erhält.

```
tweet2 <- tweet %>%
  filter(id < 51) %>%
  select(id,
         text
        )
```

In diesem möchten wir nun die Tweets finden, die das Wort *thanks* beinhalten. Die Funktion gibt einen **Boolean**-Wert zurück, also `TRUE` oder `FALSE`.

```
tweet2 <- tweet2 %>%
  mutate(includesThanks = str_detect(text,
                                     "thanks"
                                    )
         )

table(tweet2$includesThanks)
```

```
FALSE
  50
```

Kein Tweet beinhaltet also *thanks*, aber was ist mit Tweet 7?

```
tweet2$text[7]
```

```
[1] "Thanks for tweeting along with us tonight on the return of The Great British Bake Off!
You really do make the #GBBO Twitter experience pure joy. ❤️❤️ #GBBO https://t.co/PnlzGRb0Mb"
```

Kleine Denkaufgabe: Wo könnte das Problem liegen?

Das Problem liegt im großen `T`, da wir in der Anweisung nicht berücksichtigt haben, dass es uns egal ist, ob es genau die Reihenfolge ist oder ob Groß-/Kleinschreibung unberücksichtigt bleiben soll. Dazu benötigen wir die Hilfsfunktion `regex()` (Funktion für **regular expressions**), bei der wir im Argument `ignore_case = TRUE` angeben, dass es egal ist, ob die Buchstaben klein oder groß geschrieben sind, sondern nur das Muster der Reihenfolge vorhanden sein muss.

```
tweet2 <- tweet2 %>%
  mutate(includeThanks = str_detect(text,
                                    regex("thanks",
                                           ignore_case = TRUE)
                                    )
  )

table(tweet2$includeThanks)
```

```
FALSE  TRUE
   49     1
```

Nur noch einmal zur Logik von **regular expressions**: In diesem Fall möchten wir das Wort `thanks` in Tweets suchen. Dazu geben wir in `regex()` die Buchstabenreihenfolge an, die eben dieses Wort kennzeichnen, also `thanks`. Hätten wir einen Tweet mit dem Wort `thanksgiving` würde dies in diesem Fall auch als `TRUE` ausgegeben werden, da es eben auch die Buchstabenreihenfolge beinhaltet.

```
beispiel <- c("Thanks! It means a lot to me",
              "I hope I see you on Thanksgiving",
              "Maybe next time!")

str_detect(beispiel,
            regex("thanks",
                  ignore_case = TRUE)
            )
```

```
[1] TRUE TRUE FALSE
```

Um dies jetzt zu umgehen, könnte man einfügen, dass nach dem Ende der Buchstabenreihenfolge ein Leerzeichen folgen muss und davor ein Leerzeichen sein kann oder aber auch nicht (Satzanfang!).

Für Leerzeichen setzt man in **regular expressions** `\\s` in R. D.h. vor `thanks` setzen wir `\\s`. Da wir aber wollen das auch ein Wort am Satzanfang gezählt wird, setzen wir dahinter ein `?`, was bedeutet, dass die keinmal oder mehr vorkommen kann. Somit hätten wir in der **regular expression**: `\\s?thanks`. Jetzt wollen wir aber noch Wörter ausschließen, die mit `thanks` beginnen. Dazu können wir wieder ein Leerzeichen setzen. Machen wir nur das wird kein `thanks` gezählt, da im ersten Item kein Leerzeichen, sondern ein Satzzeichen folgt. Daher geben wir in der **regular expression** an, dass entweder ein Leerzeichen (`\\s`) oder ein Satzzeichen (`[[:punct:]]`) folgt. Der gesamte **regex**-Ausdruck ist also: `\\s?thanks(\\s|[[:punct:]])`. Dies setzen wir nun ein und schauen, ob die Funktion es nun richtig erkennt.

```
str_detect(beispiel,
            regex("\\s?thanks(\\s|[[:punct:]])",
                  ignore_case = TRUE)
            )
```

```
[1] TRUE FALSE FALSE
```

Wie das Beispiel verdeutlicht, kann es sehr schnell kompliziert werden mit Textdaten. Das ist nicht schlimmes, nur sollte man sich dessen bewusst sein und ziemlich konkret und genau überlegen, wie man Textdaten bearbeitet.

Will man anstatt eines **Boolean**-Wertes die Häufigkeit des Vorkommens einer Buchstabenkette erhalten, nutzt man `str_count()`. Wir könnten z. B. testen, wie oft der Name "Prue" (eine der zwei Judges) in den Tweets vorkommt.

```
tweet2 <- tweet2 %>%
  mutate(countPrue = str_count(text,
                                regex("Prue",
                                       ignore_case = TRUE)
                                )
  )

tweet2 %>%
  select(text,
         countPrue
  )
```

text
<chr>
There are two key stages in life. Before you discover Malt Loaf and everything afterwards. #GBBO https://t.co/4LSBTfhiR4
@GeekEnby We all deserve a little ray of sunshine at least once a week. 🙏 #GBBO
Well, it looks like malt loaf awareness isn't in bad shape at all! #GBBO https://t.co/4LSBTfhiR4
We bow down before all of you wonderful people who joined us last night for the return of #GBBO https://t.co/q5ebtUTMAi
When your new crush starts ghosting you... #GBBO https://t.co/XwcVnIBfCi
About last night... #GBBO https://t.co/GulcsAdnJM
Thanks for tweeting along with us tonight on the return of The Great British Bake Off! You really do make the #GBBO Twitter experience pure joy. ❤️❤️ #GBBO https://t.co/PnlzGRb0Mb
Let's get ready to crumble! It's crunch time next Tuesday as Biscuit Week hits the Bake Off Tent! #GBBO https://t.co/xGMRw30WA8
@philringsell @PrueLeith Tres Bon-o. #GBBO https://t.co/vGlYoVXeCs
@Shambom65 True. Absolutely true. #GBBO
Disclaimer: This is not a percentage in terms of flavour balance. It's how many people like Mint Choc Ch... *stares directly to camera* #GBBO
It's official! Mint Choc Chip has now been certified as being 67.6% Mint. #GBBO https://t.co/vUovHq8iqD
"Although my time in the Bake Off tent may have been short, it certainly has been very sweet." \nHere's Tom's lovely letter to all you #GBBO fans. https://t.co/JgKlfD8WDg
No expense was spared to turn the Bake Off Tent into an anti-gravity zone for tonight's show. #GBBO https://t.co/8MEF5qemLe
Mark our words... 🤔 #GBBO https://t.co/5s8w8XjVDR

1-15 of 50 rows | 1-1 of 2 columns

Previous 1 2 3 4 Next

str_extract()

Mit dieser Funktion kann ein spezifischer Textteil gesucht werden und gleichzeitig extrahiert werden. Wir erhalten dann einen neuen Vektor/Variable, die das erste Vorkommen der gesuchten Kombination enthält. Will man alle Kombinationen extrahieren, die vorliegen, nutzt man einfach `str_extract_all()`.

```
tweet2 %>%
  mutate(extractThanks = str_extract(text,
                                     regex("thanks",
                                           ignore_case = TRUE)
                                     )
  )
```

id <dbl>
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

1-15 of 50 rows | 1-1 of 6 columns

Previous 1 2 3 4 Next

Zur Erinnerung: Warum benutzen wir hier die Hilfsfunktion `regex()` ?

Will man zum Beispiel alle *mentions* in einem Tweet erhalten, nutzt man `str_extract_all()` . Diese Funktion extrahiert nicht nur das erste Vorkommen, sondern alle Vorkommen. Probieren wir es aus.

```
tweet2 <- tweet2 %>%
  mutate(mentions = str_extract(text,
                                "@"
                                )
  )

tweet2 %>%
  select(id,
         text,
         mentions
  )
```

id <dbl>
1
2

	id
	<dbl>
	3
	4
	5
	6
	7
	8
	9
	10
	11
	12
	13
	14
	15
1-15 of 50 rows 1-1 of 3 columns	Previous 1 2 3 4 Next

Denkaufgabe: Welches Problem haben wir jetzt (wieder)?

Es wird zwar das ****@*-Zeichen gefiltert, aber nicht der Text danach, mit dem wir die mention* identifizieren können.**

Wir wollen das jetzt mal probieren. Logisch benötigen wir einen Befehl, er uns sagt, dass das @-Zeichen mit den anhängenden Zeichen bis zum nächsten Leerzeichen entnommen werden soll. Bei **regex** gibt es den Unterbefehl `\w`, mit dem der nachfolgende *word character* extrahiert wird. Also mit `@\w` erhalten wir das @-Zeichen und den ersten Buchstaben der *mention*. Wenn wir nach dem `w` noch ein `+` einfügen, erhält man die nachfolgenden *word character* bis zum nächsten *non-word character*.

```
tweet <- tweet %>%
  mutate(mention = str_extract_all(text,
                                   regex("@\\w+"
                                   ),
                                   simplify = FALSE
                                   )
  )

head(tweet$mention,
      n = 50
      )
```

```
[[1]]  
character(0)
```

```
[[2]]  
[1] "@GeekEnby"
```

```
[[3]]  
character(0)
```

```
[[4]]  
character(0)
```

```
[[5]]  
character(0)
```

```
[[6]]  
character(0)
```

```
[[7]]  
character(0)
```

```
[[8]]  
character(0)
```

```
[[9]]  
[1] "@philringsell" "@PrueLeith"
```

```
[[10]]  
[1] "@Shambom65"
```

```
[[11]]  
character(0)
```

```
[[12]]  
character(0)
```

```
[[13]]  
character(0)
```

```
[[14]]  
character(0)
```

```
[[15]]  
character(0)
```

```
[[16]]  
character(0)
```

```
[[17]]  
character(0)
```

```
[[18]]  
character(0)
```

```
[[19]]
```


character(0)

[[20]]
character(0)

[[21]]
character(0)

[[22]]
character(0)

[[23]]
character(0)

[[24]]
character(0)

[[25]]
character(0)

[[26]]
character(0)

[[27]]
character(0)

[[28]]
character(0)

[[29]]
character(0)

[[30]]
[1] "@courtsthomasx"

[[31]]
character(0)

[[32]]
character(0)

[[33]]
character(0)

[[34]]
character(0)

[[35]]
character(0)

[[36]]
character(0)

[[37]]
character(0)

```
[[38]]
character(0)

[[39]]
[1] "@PrueLeith"

[[40]]
character(0)

[[41]]
character(0)

[[42]]
[1] "@psyche1701"

[[43]]
character(0)

[[44]]
[1] "@JadeForster"

[[45]]
character(0)

[[46]]
character(0)

[[47]]
character(0)

[[48]]
character(0)

[[49]]
character(0)

[[50]]
[1] "@laurajaycrochet"
```

Eine Sache muss man hier aber berücksichtigen: `str_extract()` erstellt eine Liste! Das heißt für jede Zeile im Datensatz erhalten wir eine Liste, in der dann die einzelnen *mentions* als ein Item gespeichert sind. Das ist fürs Datenhandling nicht ganz optimal.

```
tweet$mention[9][1]
```

```
[[1]]
[1] "@philringsell" "@PrueLeith"
```

Um dies wiederum aufzuheben, ist ein weiterer Schritt notwendig, so dass wir so viele Variablen für `mentions` haben, wie es maximal pro Tweet gibt. Dazu nutzen wir die Funktion `unnest_wider()` aus `tidyr`:

```
tweet <- tweet %>%
  unnest_wider(mention)
```

```
head(tweet)
```

text

<chr>

There are two key stages in life. Before you discover Malt Loaf and everything afterwards. #GBBO <https://t.co/4LSBTfhiR4>

@GeekEnby We all deserve a little ray of sunshine at least once a week. 🙏 #GBBO

Well, it looks like malt loaf awareness isn't in bad shape at all! #GBBO <https://t.co/4LSBTfhiR4>

We bow down before all of you wonderful people who joined us last night for the return of #GBBO <https://t.co/q5ebtUTMAi>

When your new crush starts ghosting you... #GBBO <https://t.co/XwcVnIBfCi>

About last night... #GBBO <https://t.co/GulcsAdnJM>

6 rows | 1-1 of 13 columns

Jetzt sollten nur noch die Spaltennamen umbenannt werden mit `rename()` (aus `dplyr`):

```
tweet <- tweet %>%
  rename("mention1" = "... 1",
         "mention2" = "... 2",
         "mention3" = "... 3",
         "mention4" = "... 4",
         "mention5" = "... 5",
         "mention6" = "... 6",
         "mention7" = "... 7",
         "mention8" = "... 8"
  )
```

```
tweet
```

text

<chr>

There are two key stages in life. Before you discover Malt Loaf and everything afterwards. #GBBO <https://t.co/4LSBTfhiR4>

@GeekEnby We all deserve a little ray of sunshine at least once a week. 🙏 #GBBO

Well, it looks like malt loaf awareness isn't in bad shape at all! #GBBO <https://t.co/4LSBTfhiR4>

We bow down before all of you wonderful people who joined us last night for the return of #GBBO <https://t.co/q5ebtUTMAi>

When your new crush starts ghosting you... #GBBO <https://t.co/XwcVnIBfCi>

About last night... #GBBO <https://t.co/GulcsAdnJM>

Thanks for tweeting along with us tonight on the return of The Great British Bake Off! You really do make the #GBBO Twitter experience pure joy. ❤️❤️ #GBBO <https://t.co/PnlzGRb0Mb>

Let's get ready to crumble! It's crunch time next Tuesday as Biscuit Week hits the Bake Off Tent! #GBBO <https://t.co/xGMRw30WA8>

@philringsell @PrueLeith Tres Bon-o. #GBBO <https://t.co/vGLYoVXeCs>

@Shambom65 True. Absolutely true. #GBBO

Disclaimer: This is not a percentage in terms of flavour balance. It's how many people like Mint Choc Ch... *stares directly to camera* #GBBO

It's official! Mint Choc Chip has now been certified as being 67.6% Mint. #GBBO <https://t.co/vUovHq8iqD>

text													▶							
<chr>																				
“Although my time in the Bake Off tent may have been short, it certainly has been very sweet.”\nHere’s Tom’s lovely letter to all you #GBBO fans. https://t.co/JgKlFD8WDg																				
No expense was spared to turn the Bake Off Tent into an anti-gravity zone for tonight’s show. #GBBO https://t.co/8MEF5qemLe																				
Mark our words... 🙄 #GBBO https://t.co/5s8w8XjVDR																				
1-15 of 862 rows 1-1 of 13 columns																				
												Previous	1	2	3	4	5	...	58	Next

[1] "@GeekEnby We all deserve a little ray of sunshine at least once a week. 🙏 #GBBO"

[2] "@philringsell @PrueLeith Tres Bon-o. #GBBO <https://t.co/vGLYoVXeCs>"

[3] "@Shambom65 True. Absolutely true. #GBBO"

[4] "@courtsthomasx Congratulations! 🥳🥳 #GBBO"

[5] "Unfortunately Prue2 is already the name of @PrueLeith's U2 tribute band. #GBBO <http://t.co/Pm97wOWQNN>"

[6] "@psyche1701 Maybe. #GBBO"

[7] "@JadeForster A treat for the whole family. 🍪 #GBBO"

[8] "@laurajaycrochet *snaps pencil* #GBBO"

[9] "@randallwrites Not in line with the currant climate of acceptance for all is it?"

[10] "@carol_ann54 We do seriously respect your integrity for sticking with the healthy option. #GBBO <https://t.co/S3aLrdef5H>"

[11] "@beckythomas_ We're happy that you're happy! #GBBO"

[12] "@Emma_Parkin Fine work. #GBBO <https://t.co/QrKF801Uy0>"

[13] "@Toxicolly Thankin' you! #GBBO"

[14] "@CabinClassBoy We would but Pet Bunny's agent is a NIGHTMARE to negotiate with... #GBBO"

[15] "@TheMizzCCB \"Mini Rolls pack a punch\" according to Paul. We've played Chocolate Chili Pepper Roulette before but this is taking it a bit far... #GBBO"

[16] "Spinoff TV series pitch: Matt Lucas sings well-known TV theme songs... in German... with Jürgen giving his honest critique afterwards. Slide into our DMs if interested, @Channel4. #GBBO"

[17] "@nickopotamus We are pre-watershed, you know. #GBBO"

[18] "@S_M_A_B_S Imagine waking up with a mullet too... #GBBO"

[19] "@HelenTWrites We have the utmost respect for your health kick. Bravo Helen, bravo! #GBBO <https://t.co/D0PeyEmucz>"

[20] "@k_louise123 We salute your assertive boundary setting skills. 🙏 #GBBO"

[21] "@CoronaSinger Bravo! #GBBO"

[22] "@laurajwilshaw_ @fretterz14 We've missed this 3000. #GBBO <https://t.co/B9rvy4cU0j>"

[23] "@laurajwilshaw_ @fretterz14 The universe is in balance again. Welcome back. #GBBO"

[24] "@kirstybaillie @RealMattLucas Noooooooo! Good luck avoiding spoilers Kirsty! #GBBO"

[25] "Are you ready to meet our twelve new bakers?\n\nThe Great British Bake Off. 8pm. @Channel4. #GBBO <https://t.co/QstjMCOJ3r>"

[26] "@lisagibbs84 Welcome back! #GBBO <https://t.co/77TanTJydD>"

[27] "All good things come to those who bake. Not long to go now!\n\n\nThe Great British Bake Off. Tonight. 8pm. @Channel4. #GBBO <https://t.co/f5AJvCzP34>"

[28] "@EllenvWallace Our first percentage answer. Also, well done! #GBBO"

[29] "@Guapaish @Channel4 Sieve and let sieve, as they* say. 🍷\n\n*Only this account. #GBBO"

[30] "@TomButler Pssst! Paul has been bread since 2017. This rabbit hole leads all the way back to the phrase \"The Male Judge\" being spoken in 2014. Off the record, of... wait... the door... who's that? You? Impossible! WaipLEASE IGNORE THIS TWEET. EVERYTHING IS FINE AND PAUL IS NOT BREAD. #GBBO"

[31] "@SeykortTheresa WHAT IS GOING ON? WHAT NEXT? HMMMN? A 50 FOOT TALL THIMBLE FULL OF GASPACHO SOUP AT THE PYRAMIDS OF GIZA? HMMMN? #GBBO <https://t.co/hzmg1WOcsA>"

[32] "@SmoggyThorpes @Channel4 Brilliant. Simply brilliant. #GBBO <https://t.co/oePYu9ui9t>"

[33] "@Guapaish @Channel4 Please, nobody grill Lesley. We are all equally entitled to our opinions."

[34] "@pawswithprose @Channel4 \"Dear God, it's beautiful.\" - Jacobim Mugatu #GBBO <http://t.co/Ec2gfsLNPD>"

[35] "@revkatebottley Amazing. #GBBO <https://t.co/huoWrjPzWD>"

[36] "@chton Skull Island. We don't like to talk about it. Or make Ice Cream Sundaes for King Kong. Not anymore. #GBBO"

[37] "@scottgyb To you too, Scott! #GBBO <https://t.co/dXqENEQSuU>"

[38] "All roads lead to the Tent of Dreams!\n\n\nThe Great British Bake Off. Tonight. 8pm. @C

hannel4. #GBBO <https://t.co/rwWuYJOxQB>"

[39] "@therealsarc True, true... #GBBO <https://t.co/jbYG7vnAJk>"

[40] "@AndreaMartIn Respect! #GBBO"

[41] "@YannyBakes ALL the points Yan, have all the points! #GBBO"

[42] "@CatherineHewins @alesterowen @thisgirlgrows Fussy... or just 100% correct? 🤔 #GBBO"

[43] "Here we dough! Join us tonight from 8pm on @Channel4 for the new series of The Great British Bake Off! #GBBO <https://t.co/Jc4kshY4xs>"

[44] "@iambrooke 🙏🙏🙏🙏 #GBBO"

[45] "@19Sunflowers We trying to concoct a Stephen \"Tiffin\" Duffy pun here but realised t he time taken to do so might not justify the risk of revealing our age. #GBBO"

[46] "@stephenhardie Be the miracle, Stephen. 🙏🙏🤔 #GBBO"

[47] "@colettem1983 We totally know but unfortunately can't share that information right no w. *Opens web browser* ... #GBBO"

[48] "@slhattersley To be honest, it's as much of a challenge as it is a question. 🤔 #GBB O"

[49] "@jonwatt876 Loving your work here Jon. Respect. 🙏 #GBBO"

[50] "@patmunn @noelfielding11 @PaulHollywood @RealMattLucas @PrueLeith You have no idea ho w much we uhhmed and aaahed about that. #GBBO"

[51] "If Bake Off were #Seinfeld who would be who?\n\nSuggestions:\n\nElaine = @noelfieldin g11 \nGeorge = @PaulHollywood \nJerry = @RealMattLucas \nKramer = @PrueLeith \n\n#GBBO <http://t.co/Nm8arY40f0>"

[52] "Prepare to enter the Tent of Dreams once more...\n\nThe Great British Bake Off return s.\nTuesday. 8pm. @Channel4. #GBBO <https://t.co/tpW8o2RScJ>"

[53] "A bake for every time of year, the official GBBO book features recipes by @PaulHollyw ood, @PrueLeith and the 2021 bakers. Pre-order your copy today: <https://t.co/IZT2iPORRq> #GBB O <https://t.co/t2IXWkhAYT>"

[54] "Craving a certain spinoff show that's a cut above? An Extra Slice returns to @Channel 4 with Jo Brand and @tomallencomey on Friday 24th September at 8pm. #GBBO #ExtraSlice <http://t.co/1qoRxEndcX>"

[55] "@DannsMrs Steady now. 🤔 #GBBO"

[56] "When you realise there's no Bake Off Pros to look forward to next Tuesday night on @C hannel4... #GBBO <https://t.co/VSXheJanfp>"

[57] "@LadyMaher Our stash has melted. #GBBO"

[58] "@veefeeley #GBBO <https://t.co/oZqfqjxBYD>"

[59] "@LynnKnight01 Phew! I mean, we admire your dedication but even we thought \"Wow. Lynn e REALLY means business here.\" #GBBO"

[60] "@WTairu Gracias! #GBBO"

[61] "@janet_faz Love the word \"giddy\"! #GBBO"

[62] "@LynnKnight01 We can ask for no more than this. Thank you. #GBBO"

[63] "@artoostacetoo 8pm! It is The Final tonight so you have [checks watch] 1 hour to cram view every episode so far. #GBBO"

[64] "It's the grand final of Bake Off: The Professionals at 8pm on @Channel4 tonight! Her e's a glimpse of the thrills, spills and skills that await... #GBBO <https://t.co/Zg1eogvZAU>"

[65] "Just three teams remain to battle it out in tonight's Final of Bake Off Pros. But who will emerge victorious? Join us at 8pm on @Channel4 to find out! #GBBO <https://t.co/Yinn6J0Qt3>"

[66] "@skippity_doo Sorry! #GBBO"

[67] "Bake your way to a banging body. #GBBO @liamcbakes <https://t.co/SEvhjdjZwnf>"

[68] "Cowpats? @tomallencomey's really put his foot in it! 🐮🐮🐮 #GBBO <https://t.co/na7LbqR70C>"

[69] "To celebrate the 25th anniversary of the @SpiceGirls, our two presenters have become one. #GBBO @tomallencomey @liamcbakes <https://t.co/gnaxtFdzcY>"

[70] "@WatchingTVnFilm No joke... #GBBO <https://t.co/PEFtmhoebu>"

[71] "The Euros and Wimbledon may be over, but it's not the end of tense Semis this summer! \nBake Off: The Professionals. Tonight. 8pm. @Channel4. #GBBO <https://t.co/nS0YhJcYCm>"

[72] "We hope @tomallencomey has been doing some catching practice ahead of tonight's Bake Off Pros Semi Final! #GBBO <https://t.co/nKcKfJGRZS>"

[73] "@ConnorT1912 There's an innuendo in this tweet? We'll have to get to the bottom of this. #GBBO"

[74] "@VicCollins89 Why thank you! Yes, all the socials. Versatility is everything. 🤔 #GBBO"

[75] "@DelightfulDame Perhaps our favourite reply, ever. #GBBO"

[76] "@Anjipowerr @PrestigePatiss Thanks for joining us! #GBBO <https://t.co/4Rz2FAigcN>"

[77] "Ah, we miss those days of youth when eating was so much simpler. And accompanied by a aeroplane noises. #GBBO @tomallencomey <https://t.co/9ShZKI8mjW>"

[78] "Boom sha-sha-shake the Cherish! 😂😂😂 #GBBO @BenoitBlin_MCA @cherish_finden <https://t.co/eqT4Cuu1Tl>"

[79] "@susikaam Thank you! #GBBO"

[80] "@AnnekaRice Thank you so much! #GBBO"

[81] "@onmoonlitshores Resistance is Choux-tile. 🤔 #GBBO"

[82] "Walking onto the dancefloor for the first time when it's safe to go boogie again. 🕺 #GBBO @tomallencomey <https://t.co/vBlLUM5zz5>"

[83] "It's almost crunch time for these five pastry chef teams. Join us at 8pm on @Channel4 to see who will make it through to the Semis. #GBBO <https://t.co/oBQ4E7NxZn>"

[84] "@veetee39 May your team enjoy glorious victory! 🏆 #GBBO"

[85] "The baking battle intensifies tonight on @Channel4 as it's Quarter Final time on Bake Off: The Professionals! #GBBO <https://t.co/BkYmLIWs6l>"

[86] "@ICUStat We're not even sure we should have asked... 🤔 #GBBO"

[87] "@haaannahrose_ We're not just here for the noisy things in life... #GBBO"

[88] "These six pastry chef duos are about to hit @Channel4 at 8pm! Who are you tipping for glory tonight? #GBBO <https://t.co/zE1VowLUd3>"

[89] "We're a little too giddy with excitement about seeing the top chefs from Heat One and Heat Two collide in Bake Off Pros at 8pm on @Channel4 tonight! #GBBO <https://t.co/WujAU1W22B>"

[90] "Thanks for joining us for Bake Off Pros! We'll hand over to @BenoitBlin_MCA for the final words of the night... #GBBO <https://t.co/fHkq1sKaAj>"

[91] "Benoit going full on Hannibal Lecter is quite something. #GBBO @BenoitBlin_MCA <https://t.co/gU90lcChjh>"

[92] "It's just not cricket! #GBBO @BenoitBlin_MCA <https://t.co/sZaa9ao9aa>"

[93] "The horror, the horror.\nWelcome to a very spooky Bake Off Pros! \n@liamcbakes @tomallencomey #GBBO <https://t.co/MRYiHb35K6>"

[94] "If you have any fingernails left to gnaw on, join us over on @Channel4 at 8pm for more knockout action in Heat Two of Bake Off Pros! #GBBO <https://t.co/i6YcJ0lAgF>"

[95] "@inyourwarmth 🙏 #GBBO"

[96] "@Preciou53411925 We'll make and eggs-ception tonight, eh? 🤔🏆🍰 #GBBO"

[97] "There's some horror in store tonight - and it's not the football! Join us for some spooky action on Bake Off Pros at 8pm on @Channel4. #GBBO <https://t.co/tLMZBMj3k0>"

[98] "When the caffeine kicks in on a Monday morning... #MondayMotivation #GBBO @tomallencomey <https://t.co/pj5cUE51cU>"

[99] "Shimmying into the weekend like... \n#GBBO @liamcbakes @tomallencomey <https://t.co/eQ2YdKzd5v>"

[100] "Trying to remember how to dance again - it's been so long! \n#GBBO @tomallencomey @liamcbakes <https://t.co/TsMLVK2xbV>"

[101] "@_holamanola @OfficialKat <https://t.co/hXWz8tAR9P>"

[102] "@NoorAlmuzaffar @OfficialKat @JamesAcastor 🤔 <https://t.co/HLqSDTxLxu>"

[103] "@OfficialKat We have WandaVision playing in the other room and our understanding of The Ship Of Theseus Experiment has gone up to 8.5.\n\nWe think. 🤔\n\n#GBBO"

[104] "@StevenJohnHamp1 You refer, of course, to our spin-off show Galaxy Que... ah. #GBBO"

[105] "Some words are simply too much of a mouthful. #GBBO @tomallencomey <https://t.co/oLRxRdpLsJ>"

[106] "@MrsCruz1999 https://t.co/1mkxELs2F9"
[107] "@sarahmas Sincere apologies. 🙏 #GBBO"
[108] "@amrfarid140 @areej_abuali We're not sure we can do anything about this either! #GBB
O"
[109] "The worlds of Bake Off and Doctor Who meet at last! Is this... canon? #GBBO @LiamcBakes
@tomallencomey https://t.co/CJlLnM9IwV"
[110] "@susikaam @Channel4 We're just grateful you're finding the EXTRA TIME (eh? EH?!?) to
watch our show, Sue! #GBBO #NobelPrizeForWordPlayCandidate"
[111] "If you hear any cheering after 8pm tonight, it's for the baking brilliance of these f
ive teams of pastry chefs! Bake Off Pros, coming up shortly on @Channel4... #GBBO https://t.c
o/rDCLvSSMyf"
[112] "@nualasalter Something must fill the void and that something is competitive bakery. #
GBBO"
[113] "What your baking brain tells you if you have plans to watch anything else other than
Bake Off Pros at 8pm on @Channel4 tonight... #GBBO https://t.co/nG1ltWJZgI"
[114] "As Heat Two Week Two of Bake Off Pros hits @Channel4 tonight, here's a reminder of th
e remaining teams battling it out to impress the judges... #GBBO https://t.co/zOBHOPFDtc"
[115] "Bake Off: The Professionals\nTonight. 8pm. @Channel4. \n(Also available afterwards on
@All4 😊😊)"
[116] "@coolAngieWalton @SwindyMiller White chocolate is always better than no chocolate tho
ugh.... #GBBO #Truth"
[117] "@coolAngieWalton No comment. 😊 #GBBO"
[118] "@LaLOctavia @Ariadne_Griffin We have an incredible production team. Honestly. Fantast
ic. All the work and that goes into making our shows... they're incredible. #GBBO"
[119] "Dear @mgmstudios, please consider this man to replace Daniel Craig as your next James
Bond @007. #GBBO @BenoitBlin_MCA https://t.co/Ahk9bMbvnC"
[120] "@CelebrantJayne \"We've been bake-to-bake, sure. I will not hesitate. Not for a secon
d.\" #GBBO"
[121] "Dear @GaryLineker - one of our pastry chefs on Bake Off: The Professionals is named a
fter you! 🍪🍪 #BakeOfTheNet #GBBO https://t.co/VqUcoGsY2Q"
[122] "@thyroidlady @Channel4 Very kind of you to say so. Thrilled you enjoy our show! #GBB
O"
[123] "Coming up at 8pm on @Channel4, Heat Two and these fresh faces whisking it all to impr
ess the judges... #GBBO https://t.co/qurGrHQdPh"
[124] "Constructively passing the time stuffing my face until Heat Two of Bake Off: The Prof
essionals kicks off at 8pm on @Channel4. #GBBO https://t.co/b2FhoDrcY5"
[125] "Heat Two of Bake Off Pros kicks off tonight! \nCan any of the new chefs' creations pu
t a smile on the faces of @BenoitBlin_MCA and @Cherish_Finden? #GBBO https://t.co/mTclg3pZC
j"
[126] "@simonheath81 @cherish_finden @duranduran Sorry, I don't want your loaf."
[127] "@simonheath81 @cherish_finden @duranduran You've gone choux far this time."
[128] "@simonheath81 @cherish_finden @duranduran Yes. The union of the bake is on the climb.
#GBBO"
[129] "@lozzzknight Glad to be of service! #GBBO"
[130] "@laurabarr38 We're so glad you got that reference! Imagine if Tick Tock was on TikTo
k... #GBBO https://t.co/7GDA8FjtCy"
[131] "Benoit says it best when he says nothing at all. #GBBO @BenoitBlin_MCA https://t.co/F
OTeqI26p0"
[132] "@WendyTh50990762 It's only going one of two ways, isn't it? #GBBO"
[133] "Terminator: Rise Of The Liams. #GBBO @Liamcbakes https://t.co/pUYmBQ9hjV"
[134] "\"Stunning\" is the verdict from @BenoitBlin_MCA on Julien & Elise's Chocolate and
Pink Grapefruit Cake. #GBBO https://t.co/fkbXN7BTsQ"
[135] "@jambosn2 Our pleasure! Hope you enjoy the show tonight! #GBBO"
[136] "We're with you 100% @BenoitBlin_MCA! \n💎💎💎💎💎💎💎💎💎💎💎💎💎💎💎💎💎💎\n😄😄😄😄😄😄😄😄😄😄😄😄😄😄😄😄😄😄\n#GBBO https://t.co/8uOrN84r4D"

[137] "The greatest crossover event of all time. @Cherish_Finden @DuranDuran #GBBO <https://t.co/p99vYAA11V>"

[138] "That 'Weird Science' remake is coming along well. #GBBO @tomallencomedy @liamcbakes <https://t.co/X92HRA1z41>"

[139] "Your default setting for tonight's Chocolate Week.\nBake Off: The Professionals. 8pm. @Channel4. #GBBO <https://t.co/BJaa2MYXM8>"

[140] "Join @Liamcbakes, @tomallencomedy, @BenoitBlin_MCA and @Cherish_Finden for an hour of telly guaranteed to make your mouth water and make your heart boom boom a little bit faster.\nBake Off: The Professionals. Chocolate Week. Tonight. 8pm. @Channel4. #GBBO <https://t.co/l4A05ITyT3>"

[141] "Here are the remaining pastry duos battling it out in Heat One Week Three of Bake Off: The Professionals. And there's a little secret in store for them too. Join us at 8pm tonight on @Channel4 to find out! #GBBO <https://t.co/SGShKtSpIR>"

[142] "All good things come in twos. Join @tomallencomedy and @liamcbakes for Heat One Week Three of Bake Off: The Professionals tonight at 8pm on @Channel4. #GBBO <https://t.co/YaBjXolckI>"

[143] "@Ianhwatkins Tell me that you are the only bun. The only bun for me ..."

[144] "@Ianhwatkins If you do make a big cake, please make sure there aren't any wasted tiers ... 🙄🙄"

[145] "@Ianhwatkins The love has scone ... but it's got a hold of our tart! #GBBO"

[146] "When well intended mushroom puns turn out a bit shiitake... and we thought @tomallencomedy was such a fun-ghi! #GBBO <https://t.co/yDqx0LnBZI>"

[147] "When the muscle memory kicks in during an @OfficialSteps song next time you're allowed on the dance floor... #GBBO @Ianhwatkins <https://t.co/tcwFtuS50V>"

[148] "@PatchworkSteve Team work makes the dream work... #GBBO <https://t.co/L7uPLjrRbp>"

[149] "@WatchingTVnFilm Bake Off moves pretty fast... #GBBO <https://t.co/l0b3AUznN9>"

[150] "🌶🌶🌶 #GBBO @Liamcbakes <https://t.co/cYAHczx9uf>"

[151] "@ylovesgok @Channel4 You say \"toadstool\", we say \"mushroom\" ... that really doesn't work does it? <https://t.co/7a7zDwx96w>"

[152] "🤖🤖🤖🤖 #GBBO @Cherish_Finden <https://t.co/iYDEbeh0Fo>"

[153] "The most feared sight in baking - @Cherish_Finden with her ruler out. #GBBO <https://t.co/D0319vP9Gv>"

[154] "@CMNatic Thank you. The key to posting for so many platforms is being versatile. #GBBO"

[155] "Now we may be wrong, but we get the impression @Cherish_Finden may have enjoyed that. #GBBO <https://t.co/GMmrHG31KF>"

[156] "@Apshe1 🤗 #GBBO"

[157] "Me during every #GBBO ad break. @tomallencomedy <https://t.co/pQs7FXiFCp>"

[158] "@BurgessMoon A damn fine spot on your part there... #GBBO <https://t.co/AI4KpYkRco>"

[159] "Iconic. ❤️❤️\n@liamcbakes displays his superpowers and @tomallencomedy shows us his Fanny Cradock. #GBBO <https://t.co/79kdGQ40tb>"

[160] "It's time to calmly head towards your tellies - Bake Off Pros is coming up at 8pm on @Channel4! #GBBO <https://t.co/vEbHytPHys>"

[161] "Who faces the chop tonight? Here are the remaining pastry chef duos left to battle it out in tonight's Bake Off: The Professionals at 8pm on @Channel4. #GBBO <https://t.co/Z4xWc2NxnB>"

[162] "When you check what's on later and spot Bake Off: The Professionals in the schedules at 8pm tonight on @Channel4. #GBBO @Liamcbakes <https://t.co/a3K8abMeyC>"

[163] "Hell hath no fury like the underwhelmed palates of Benoit and Cherish. \nBake Off: The Professionals. Tonight. 8pm. @Channel4. #GBBO <https://t.co/0P0oM3K0bN>"

[164] "@robertpreseau Cookies pretty much covers it, perhaps. Very similar. #GBBO"

[165] "@GThomas1970 @BenoitBlin_MCA Great minds. 🙏 #GBBO <https://t.co/PSHq7ZXb6f>"

[166] "@StaceyAllton Have you RTd for cake? #GBBO"

[167] "@colettem1983 So that's an RT then? 🙄 #GBBO"

[168] "@boomboxgaz Good God man! We're bakers, not confectionary tax inspectors! #GBBO <http>

s://t.co/OMFBR0pxrt"

[169] "The struggle is real. #GBBO @tomallencomedy <https://t.co/fylCGfhrWK>"

[170] "@HogarthMatty #GBBO <https://t.co/rpz0fMRxg2>"

[171] "There's nothing on television quite like Benoit Blin reacting to a great dessert. #GBBO @BenoitBlin_MCA <https://t.co/SvsBJXJflb>"

[172] "In a good way, right @Gemified? #GBBO <https://t.co/HHqUrPJJeRo>"

[173] "@NoWireHangers 🤔 #GBBO <https://t.co/pPvHWVI4mp>"

[174] "@TeachingTheVoid Brilliant! Hope you enjoy the show! 🙏 #GBBO"

[175] "Things That Make A Baker Nervous #1\n\n@Cherish_Finden rolling up her sleeves before judging your work\n#GBBO <https://t.co/0XhVAA9jw0>"

[176] "Jaffa cakes are so much more than just food. #GBBO @LiamcBakes @tomallencomedy <https://t.co/NKQuURcYqw>"

[177] "@katsitter Sorry Kath, hope we're at least a good thing to enjoy in the evening. 🍷 #GBBO"

[178] "@Suedoswede Very valid point. #GBBO <https://t.co/AMr5y1CQ5B>"

[179] "No pressure, chefs. No pressure at all. #GBBO @Cherish_Finden <https://t.co/nxCt5yjcLn>"

[180] "The motivational quote we all need right now. #GBBO @BenoitBlin_MCA <https://t.co/EijaSMcpL5>"

[181] "Cherish weighs in on The Great Jaffa Cake Debate! #GBBO @Cherish_Finden <https://t.co/al954WHkd1>"

[182] "Every cloud...\nBake Off: The Professionals. Tonight. 8pm. @Channel4. #GBBO <https://t.co/eHkCBnDPo3>"

[183] "Urging the clock to get a move on so #GBBO can be back on our tellies sooner!\nBake Off: The Professionals. Tonight. 8pm @Channel4. <https://t.co/3rk7hs2W1e>"

[184] "MOUTHS. WILL. WATER.\nGet ready for a new series of Bake Off: The Professionals. Starts tonight at 8pm on @Channel4. #GBBO\n@BenoitBlin_MCA @Cherish_Finden @LiamcBakes @tomallencomedy <https://t.co/Gb9XRJzbrI>"

[185] "There ain't no pastry like a Bake Off Pros pastry. \nWelcome this awesome foursome back onto your tellies as a new series of Bake Off: The Professionals begins tonight at 8pm on @Channel4! #GBBO @BenoitBlin_MCA @LiamcBakes, @TomAllenComedy @Cherish_Finden <https://t.co/sMBNwU3y4Q>"

[186] "Here are the pastry chef duos battling it out to impress @BenoitBlin_MCA and @Cherish_Finden in Heat One of Bake Off: The Professionals tonight! #GBBO <https://t.co/PeLwtmwzXN>"

[187] "It's got to be perfect. Join us tonight for a new series of Bake Off: The Professionals with @BenoitBlin_MCA, @LiamcBakes, @TomAllenComedy and @Cherish_Finden! #GBBO <https://t.co/PamvTHcRsv>"

[188] "The proof of the pudding... is in the palates of @BenoitBlin_MCA and @Cherish_Finden!\nBake Off: The Professionals. Starts Tuesday. 8pm. @Channel 4. #GBBO <https://t.co/vk9i9V0FzC>"

[189] "Just pudding it out there - a new series of Bake Off: The Professionals with @TomAllenComedy and @LiamcBakes is almost upon us! Join us this Tuesday at 8pm on @Channel4. #GBBO <https://t.co/agbbNDOIsP>"

[190] "When there's a new series of Bake Off Pros to look forward to next week on @Channel4. #FridayFeeling #GBBO <https://t.co/qoqnNDCX8M>"

[191] "Join @LiamcBakes and @TomAllenComedy for a tasty treat this Tuesday - a new series of Bake Off: The Professionals on @Channel4! #GBBO <https://t.co/mJSGHq5uRq>"

[192] "Toilets! Monkeys! Snow Leopards! Rasta Carrots! Little Mix Concerts!\n\nHere's our countdown of the most memorable bakes from The Great Stand Up To Cancer Bake Off 2021! Featuring @KSI @Baddiel @TomAllenComedy @ReeceShearsmith @LittleMix @DizzeeRascal @AnneMarie #GBBO <https://t.co/j166g1NKRa>"

[193] "Holy moly! Can our new batch of pastry chefs impress the judges in the new series of Bake Off: The Professionals? Start licking your lips as the journey begins this Tuesday at 8pm on @Channel4. #GBBO\n\n@BenoitBlin_MCA @Cherish_Finden @LiamcBakes @tomallencomedy <https://t.co/p3CIgJLaHT>"

[194] "Fancy a laugh or ten? Here's our countdown of the funniest moments from The Great Sta

nd Up To Cancer Bake Off 2021!\n\nnFeaturing @LittleMix @KSI @NadineCoyleNow @Kathbum @Reeces
hearsmith @AnnekaRice @TomAllenComedy @JohnBishop100 #GBBO <https://t.co/cOQzatfJ5C>"

[195] "Meet the pastry chefs aiming to amaze the palates of our judges on the new series of
Bake Off: The Professionals! Join @BenoitBlin_MCA, @Cherish_Finden, @LiamCBakes and @tomalle
ncomedy next Tuesday at 8pm on @Channel4 as the quest for pâtisserie perfection begins! #GBB
O <https://t.co/DPCXDmH7Vj>"

[196] "We expect nothing less than perfection.\n\nJoin Benoit, Cherish, Liam and Tom for a n
ew series of Bake Off: The Professionals on Tuesday May 25th at 8pm on @Channel4. #GBBO <https://t.co/yIIv7UoKFK>"

[197] "Fresh from her visit to the Bake Off Tent, we challenged @AnnekaRice to name as many
bakes as possible in 30 seconds! Here's how she got scone... #GBBO <https://t.co/tNrvx7iDe6>"

[198] "The showstoppers that dreams are made of. Here's what @AdeAdepitan, @AnnekaRice, @Nad
ineCoyleNow and @JohnBishop100 created when asked to make a chocolate cake representing some
thing from their bucket list. #GBBO <https://t.co/rm6bQfkoBq>"

[199] "Genius #WednesdayWisdom from @AdeAdepitan! #GBBO <https://t.co/qEGoBL4LfN>"

[200] "@AnnekaRice We can't think of any combo you wouldn't wear well.❤️ Thanks for being suc
h a fantastic part of our show and helping raise awareness for @SU2CUK. 🙏 #GBBO"

[201] "Do it! Here are a few words for you from @AdeAdepitan.\nTimes are tough right now, bu
t any money you're able to give will really make a difference and be massively appreciated.
Together we can beat cancer! 🍌 <https://t.co/cL1kNdfsHG> #GBBO @SU2CUK <https://t.co/GCi5J83q0R>"

[202] "Nobody sees the dough until her heart says so! Nadine Coyle definitely created a whol
e lotta history in the Bake Off tent. ❤️ #GBBO @NadineCoyleNow <https://t.co/u7Bt08zqBY>"

[203] "Who needs a jumpsuit when you can have a Star Baker apron? Congratulations to @Anneka
Rice - another challenge successfully won! #GBBO <https://t.co/WuGJZ4Q7Tl>"

[204] "Bake me up! Here's @NadineCoyleNow's Trip To Venice chocolate cake showstopper. #GBBO
<https://t.co/rHz0BdzhPL>"

[205] "Hoop dreams are made of these. Feast your eyes on @AdeAdepitan's Chocolate Orange Bas
ketball Stadium! 🏀🏀 #GBBO <https://t.co/FHii5Cap0r>"

[206] "Game, set and ganache! Here's @JohnBishop100's ace 'Winning At Tennis' Cake served up
for your enjoyment. 🎾 #GBBO <https://t.co/oo1CiclZmi>"

[207] "Get to the chopper! Here's @AnnekaRice's Helicopter Flight Showstopper chocolate cake
- a treasure worth hunting for. #GBBO <https://t.co/nBBYvUqISc>"

[208] "@SiobhanMoore90 Perhaps if one our episodes ever gets to a tie-breaker situation... #
GBBO"

[209] "This is no good advice from Nadine Coyle. This is great advice! \nEvery donation to S
tand Up To Cancer matters.\n<https://t.co/cL1kNdfsHG> #GBBO @SU2CUK @NadineCoyleNow <https://t.co/ehHSdtB48p>"

[210] "What would life have been like for John Bishop on Brookside? We ask all the important
questions on Bake Off! #GBBO @realmattlucas @JohnBishop100 <https://t.co/vyqhdw7iOH>"

[211] "It's not easy being a camera operator in the Bake Off Tent. Especially when Nadine Co
yle is around. 🤪🤪🤪 #GBBO @NadineCoyleNow <https://t.co/oRM6PyHGHb>"

[212] "Understatement of the year from John Bishop! #GBBO @JohnBishop100 <https://t.co/e2obtwaoFc>"

[213] "And now for a message to you from @JohnBishop100.\n\nIf you're in a position to dip i
nto your pockets and help out, please know that you really are making a difference. <https://t.co/cL1kNdfsHG> #GBBO @SU2CUK <https://t.co/c4X3ZZKi74>"

[214] "When caramelisation goes wrong. #GBBO @realmattlucas @JohnBishop100 <https://t.co/RQP4Rdm1u2>"

[215] "Of all the episodes to forget the flour, @JohnBishop10! 🤔 How? HOW? #GBBO <https://t.co/BzAEv5BIBH>"

[216] "Picture of choux. @AnnekaRice finds a new way to win over @PaulHollywood in the Bake
Off Tent! #GBBO <https://t.co/ibr8w2fEBm>"

[217] "All who have made @NadineCoyleNow's \"I messed up\" face during a work conference cal
l say \"AYE.\" #GBBO <https://t.co/ObUaPlxKiG>"

[218] "Now it's time for @AnnekaRice to challenge you!\n\nAny money you're able to donate to Stand Up To Cancer will make a difference and help to get life-saving research back on track. <https://t.co/cL1kNdfsHG> #GBBO @SU2CUK <https://t.co/461frGZKEN>"

[219] "We challenged @AnnekaRice to destroy our carpet. Mission accomplished! #GBBO <https://t.co/J0aiDfySDQ>"

[220] "Excellent life strategy from @AdeAdepitan! 🤔🤔🤔🤔 #GBBO <https://t.co/y3yhmJHGUs>"

[221] "(Very) reluctant dropping consistency from @JohnBishop100. #GBBO <https://t.co/GcD5hVlwRj>"

[222] "@RyanJL74 @NadineCoyleNow It's beautiful 'cause you... I'll stop right there!"

[223] "1st rule of @NadineCoyleNow Is On Bake Off Club? #GBBO <https://t.co/k7CA8JTF9P>"

[224] "Seeing the Bake Off Tent is some kind of miracle for Nadine Coyle. #GBBO @NadineCoyleNow <https://t.co/o7kdFzxOaN>"

[225] "Wish this fabulous foursome luck in the Bake Off Tent - they're gonna knead it! \nTonight's lineup: \n⭐ @NadineCoyleNow \n⭐ @AnnekaRice \n⭐ @AdeAdepitan \n⭐ @JohnBishop100\n#GBBO <https://t.co/X51gKfyIk9>"

[226] "What do you want to hear @NadineCoyleNow say tonight? A) \"Gimme the tin\" OR B) \"Flour\"?\n#GBBO <https://t.co/EMwD8S4445>"

[227] "Not long to go until you see their dough! Stay tuned to @Channel4 to see how @JohnBishop100, @AnnekaRice, @AdeAdepitan and @NadineCoyleNow cope with life in the tent! \nThe Great Stand Up To Cancer Bake Off. Tonight. 8pm. #GBBO <https://t.co/6fWaJWeybU>"

[228] "Baking a difference in the tent tonight are:\n⭐ @NadineCoyleNow \n⭐ @AnnekaRice \n⭐ @AdeAdepitan \n⭐ @JohnBishop100\nThe Great Stand Up To Cancer Bake Off. Tonight. 8pm. @Channel4. #GBBO <https://t.co/C0jLV912vv>"

[229] "@exubesq @GBBOthoughts The pilot really was a very different format. #GBBO"

[230] "The Tent Of Dreams awaits @NadineCoyleNow, @AdeAdepitan, @AnnekaRice and @JohnBishop100... but will they serve our judges the stuff of nightmares? \nThe Great Stand Up To Cancer Bake Off. Tonight. 8pm. @Channel4. #GBBO <https://t.co/uq8x56Z9Zt>"

[231] "You can't mistake their pie-ology. \nThe Great Stand Up To Cancer Bake Off. Tonight. 8pm. @Channel4. #GBBO <https://t.co/uHFX3YhLXf>"

[232] "Soon to star in Doctor Who, @JohnBishop100 is about to discover that the Bake Off Tent is indeed bigger on the inside than on the outside. The magic of television.\nThe Great Stand Up To Cancer Bake Off. Tonight. 8pm. @Channel4. #GBBO <https://t.co/uylvtyBHEJB>"

[233] "Swapping jumpsuit for apron and chopper for chopping board, we challenge @AnnekaRice to win the Star Baker award! \nThe Great Stand Up To Cancer Bake Off. Tonight. 8pm. @Channel4. #GBBO <https://t.co/xv5R84KxB4>"

[234] "Will @AdeAdepitan earn another MBE, but this time for Master Baker Extraordinaire? We'll soon find out!\nThe Great Stand Up To Cancer Bake Off. Tonight. 8pm. @Channel4. #GBBO <https://t.co/MFwhpe2Nf1>"

[235] "Nobody sees the dough until her heart says so. Loaf machine @NadineCoyleNow hopes to serve up something kinda choux, but what will the judges say? \nThe Great Stand Up To Cancer Bake Off. Tonight. 8pm. @Channel4. #GBBO <https://t.co/nQCuVNELRg>"

[236] "Take a look at who we're serving up on tonight's The Great Stand Up To Cancer Bake Off! \n🍰 Lineup 🍰\n⭐ @NadineCoyleNow \n⭐ @AnnekaRice \n⭐ @AdeAdepitan \n⭐ @JohnBishop100\n#GBBO <https://t.co/rwt2n79oI8>"

[237] "It's time for these loaf machines to bake a difference. Hitting the tent for the last in the series of The Great Stand Up To Cancer Bake Off are @JohnBishop100, @AdeAdepitan, @AnnekaRice and @NadineCoyleNow! #GBBO <https://t.co/f1rQzpfbrx>"

[238] "It didn't take Black Magic for @LittleMix's Jade to fall under the spell of @PaulHollywood's eyes on Bake Off! 🤔🤔🤔🤔 #GBBO <https://t.co/8LU1fByziR>"

[239] "Pud you rather? It's time to ask all the hard-hitting questions to this week's bakers @Kathbum, @KSI, Stacey Dooley and @LittleMix's Jade Thirlwall! #GBBO <https://t.co/naIZYLTDEt>"

[240] "Who will be a loaf machine in next week's The Great Stand Up To Cancer Bake Off? Join us next Tuesday as @NadineCoyleNow, @AdeAdepitan, @AnnekaRice and @JohnBishop100 hit the Bake Off Tent! #GBBO <https://t.co/Zh889Pdd8a>"

[241] "Times are tough, but any money you're able to donate to Stand Up To Cancer will make a difference and help to get life-saving research back on track. <https://t.co/cL1kNdfsHG> #GBBO @SU2CUK @realmattlucas <https://t.co/EmmTWld0SH>"

[242] "There's no shortage of drama with Jade Thirlwall from @LittleMix in the Bake Off Tent! #GBBO <https://t.co/5vjULXMK6T>"

[243] "Out of all the bakers who have set foot in our tent, none have had a journey quite like JJ/@KSI. 🤔🤔🤔 #GBBO <https://t.co/QQ8lyDaGV2>"

[244] "Her wings were made for pie! We salute @LittleMix's Jade Thirlwall for her fine efforts tonight - great baking is clearly in her DNA. 🙌🙌🙌 #GBBO <https://t.co/64lbqq2RMZ>"

[245] "Dear @GloriaEstefan, here's a lovely message from Stacey Dooley. She has just created you in biscuit form on Bake Off. ❤️ #GBBO <https://t.co/ww58a7otLc>"

[246] "All's well that ends well. Congratulations, JJ! #GBBO @KSI <https://t.co/toZv89lqIO>"

[247] "Whether you call him JJ or @KSI, one thing's for certain - his 'Can't Live Without My Room' biscuit showstopper has landed a knockout blow! #GBBO <https://t.co/RyjuRZNA47>"

[248] "From Glory Days to a glorious glaze - here's Jade Thirlwall's 'Can't Live Without a @LittleMix Concert' biscuit showstopper. #GBBO <https://t.co/OuVSQLC0kL>"

[249] "Bring on the Pie-ami Sound Machine! Here's Stacey Dooley's 'Can't Live Without Mornin'g Tea and @GloriaEstefan' biscuit showstopper. #GBBO <https://t.co/BtI6LwmMSX>"

[250] "Here's Katherine Ryan's truly pawsome 'Can't Live Without Dogs' biscuit scene. The ultimate canine showstopper! 🐾 #GBBO @kathbum <https://t.co/t5LjZdSrIv>"

[251] "There's no time like now. Listen to @KSI and bake a difference. Times are tough, but any money you're able to give will be massively appreciated. Together we can beat cancer! 🙏 <https://t.co/cL1kNdfsHG> #GBBO @SU2CUK <https://t.co/McID95pukG>"

[252] "That Matt Lucas is a rascal! 🤔🤔🤔 #GBBO @realmattlucas <https://t.co/B18mnAi0R5>"

[253] "Salute! ❤️❤️❤️❤️ #GBBO @LittleMix <https://t.co/HIUMmtQhq8>"

[254] "Mental Note: Send @LittleMix the dry cleaning bill for our carpet. 🤔🤔🤔 #GBBO <https://t.co/Ot3Af5p2Q3>"

[255] "Rhythm is gonna get choux! It was only a matter of time before we had a @GloriaEstefan showstopper on Bake Off... #GBBO <https://t.co/A80cj8PwxI>"

[256] "Pure baking innovation from Katherine Ryan worthy of a Hollywood Handshake. 🙌🙌🙌 #GBBO @kathbum <https://t.co/juHljN0toR>"

[257] "Therapy. #GBBO @KSI <https://t.co/ynrf3M32Xz>"

[258] "Taxi for JJ! 🤔🤔🤔 #GBBO @KSI <https://t.co/wUye0VN6UM>"

[259] "Every donation to Stand Up To Cancer speeds up life-saving research. If you're in a position to dip into your pockets and help out, please know that you really are making a difference. <https://t.co/cL1kNdfsHG> #GBBO @SU2CUK @kathbum <https://t.co/hyuQcZZ5qt>"

[260] "Children shouldn't have to go through this. Families shouldn't have to go through this. Our hearts are with you, brave Isla. ❤️❤️❤️\nEvery penny of the money you donate to Stand Up To Cancer goes towards funding cancer research. <https://t.co/cL1kNdfsHG> #GBBO @SU2CUK <https://t.co/v22Dgu4La4>"

[261] "“This might have been a bad decision” 🤔🤔🤔 #GBBO @KSI <https://t.co/9omLeUaa85>"

[262] "Inspiration for latticing a pie can come from unexpected places. #GBBO @kathbum <https://t.co/dQ3a05443d>"

[263] "@LittleMix @SU2C Thank you so much! #GBBO"

An dieser Stelle nicht verzweifeln! **Regular expressions** scheinen am Anfang sehr kompliziert, aber das täuscht! Man muss sich einfach nur ein bisschen einarbeiten und mit **regex** üben.

str_replace()

Oftmals müssen Textdaten vor der eigentlichen Textanalyse bereinigt werden. In Tweets hat man oftmals unnötige Umbrüche die im Tweet-Text mit \n gekennzeichnet werden. Oder man findet doppelte Leerzeichen. Mit `str_replace()` kann man so etwas entfernen. Auch hier verwenden wir wieder **regex**, um die Muster zu identifizieren.

Denn wir wollen wir die Textanalyse nur reinen Text vorliegen haben. Zuerst entfernen wir die Umbrüche (\n) und dann die mehrfachen Leerzeichen.

```
tweet$text[13]
```

```
[1] "\"Although my time in the Bake Off tent may have been short, it certainly has been very  
sweet.\"\\nHere's Tom's lovely letter to all you #GBBO fans. https://t.co/JgKlfD8WDg"
```

```
tweet <- tweet %>%  
  mutate(text = str_replace_all(text,  
                                regex("\\n",  
                                      ignore_case = TRUE  
                                ),  
          text = str_replace_all(text,  
                                regex(" +"),  
                                " "  
          )  
  )  
  
tweet$text[13]
```

```
[1] "\"Although my time in the Bake Off tent may have been short, it certainly has been very  
sweet.\"Here's Tom's lovely letter to all you #GBBO fans. https://t.co/JgKlfD8WDg"
```

Für die reine Textanalyse entfernt man oftmals auch die URLs, die Hashtags sowie die *mentions*:

```
tweet$text[9]
```

```
[1] "@philringsell @PrueLeith Tres Bon-o. #GBBO https://t.co/vGlyoVXeCs"
```

```
tweet <- tweet %>%
  mutate(cleanText = str_replace_all(text,
                                     regex("http\\S+\\s*"),
                                     ""),
         cleanText = str_replace_all(cleanText,
                                     regex("#\\w+"),
                                     ""),
         cleanText = str_replace_all(cleanText,
                                     regex("@\\w+"),
                                     ""),
         cleanText = str_replace_all(cleanText,
                                     regex(" +"),
                                     " "))
  )

tweet$cleanText[9]
```

```
[1] " Tres Bon-o. "
```

Jetzt haben wir noch am Beginn des Strings und am Ende ein Leerzeichen, aber auch dieses können wir entfernen:

```
tweet <- tweet %>%
  mutate(cleanText = str_replace(cleanText,
                                  regex("^ "),
                                  ""),
         cleanText = str_replace(cleanText,
                                  regex(" $"),
                                  ""))
  )

tweet$cleanText[9]
```


```
[1] "Tres Bon-o."
```

Jetzt ist der Text bereinigt und wir haben in der Variable `cleanText` den reinen Text, den man dann in weiteren Anwendungen analysieren kann.

Das war's

Das waren die Grundfunktionen aus dem Paket mit `stringr`.

Die Übungsaufgaben finden sich in der RStudio Cloud im Projekt `task stringr`.

Copyright:  (<http://creativecommons.org/licenses/by-nc-sa/4.0/>)
B. Philipp Kleer, Justus-Liebig-Universität Gießen.